### Udacity Data Analyst Nanodegree: Data Wrangling Project

- Data wrangling, which consists of:
  - Gathering data
  - Assessing data
  - Cleaning data
- Storing, analyzing, and visualizing your wrangled data
- Reporting on 1) your data wrangling efforts and 2) your data analyses and visualizations

# Gathering Data

I gather the data from three sources in this project:

- twitter_archive_enhanced.csv
- image_predictions.tsv
- tweet_json.txt

# Assessing Data

I assess the data from these perspectives:

- df.info()
- df.shape
- df.describe()
- df.isnull().sum()
- df.duplicated().sum()
- df.nunique()
- df.head(2)

# Cleaning Data

I clean the data based on these 10 issues—8 quality issues and 2 tidiness issues:

Quality issues:

1. df_twitter has has invalid entries in **name** column
2. df_twitter hsas 181 columns of retweeted data that need to be dropped
3. Correct df_twitter.timestamp datatype to **DateTime**
4. We only need tweets with image and should drop the one without image
5. Drop unrelated columns
6. Tweet_id should be string data type
7. Drop any rows whose rate_denominator is not 10
8. So many dogs missing name and cannot be corrected due to limited information

Tidiness issues:

1. Merge dog stages columns into one single columns
2. Merge 3 tables into one single table

# Storing Data

I Store the clean DataFrame(s) in a CSV file with the main one named ‘twitter_archive_master.csv’

# Analyzing, and Visualizing Data

1. I analyze the most popular tweet picture and the highest retweet picture and it terms of they are the same picture
2. I also analyze and print out the picture with the lowest rating based on numerator and denominator
3. I visualize the number of tweets based on months
4. I visualize the relationship between favorite count and retweet count using scatter plot
5. Finally, I write everything I found into this report for the sake of readers.